Comparing Contextual Embeddings for Semantic Textual Similarity in Portuguese

José E. Andrade Junior[1,2], Jonathan Cardoso-Silva[3,4], and Leonardo C. T. Bezerra[1]

[1] IMD, Universidade Federal do Rio Grande do Norte, Natal, RN, Brazil
[2] iFood, Osasco, SP, Brazil
jose.andrade.099@ufrn.edu.br, leobezerra@imd.ufrn.br
[3] Data Science Brigade, Porto Alegre, RS, Brazil
[4] London School of Economics and Political Science, London, UK
jonathan.car.silva@gmail.com

Abstract. Semantic textual similarity (STS) measures how semantically similar two sentences are. In the context of the Portuguese language, STS literature is still incipient but includes important initiatives like the ASSIN and ASSIN 2 shared tasks. The state-of-the-art for those datasets is a contextual embedding produced by a Portuguese pre-trained and fine-tuned BERT model. In this work, we investigate the application of Sentence-BERT (SBERT) contextual embeddings to these datasets. Compared to BERT, SBERT is a more computationally efficient approach, enabling its application to scalable unsupervised learning problems. Given the absence of SBERT models pre-trained in Portuguese and the computational cost for such training, we adopt multilingual models and also fine-tune them for Portuguese. Results showed that SBERT embeddings were competitive especially after fine-tuning, numerically surpassing the results of BERT on ASSIN 2 and the results observed during the shared tasks for all datasets considered.

Keywords: Deep learning · Natural language processing · Semantic textual similarity. Word embeddings.

1 Introduction

Semantic textual similarity (STS) is the task of measuring how semantically similar a pair of sentences is [11]. The importance of this task to the natural language processing (NLP) field is endorsed by the creation of STS shared tasks, such as the ones proposed by the International Workshop on Semantic Evaluation (*SemEval* [12, 5]). Shared tasks like SemEval led to great advancements in STS, but there is still a reduced number of studies in the context of the Portuguese language. To foster this research, the ASSIN [9] and ASSIN 2 [15] workshops hosted shared tasks for STS and natural language inference (NLI) in Portuguese. The models developed during ASSIN 2 used more recent NLP approaches, including contextual embeddings like BERT [7]. Recent works addressing these datasets have since been continuously proposed, with the state-of-the-art being a BERT model pre-trained in Portuguese fine-tuned for STS [8].

Though competitive, training BERT models on STS demands that all pairs of sentences be used as input to the network. This approach can cause a massive computational overhead, even for a moderate-size corpus. To address this problem, Reimers and Gurevych [17] proposed Sentence-BERT (SBERT), a siamese architecture with shared weights that can reduce the cost of BERT models. Since its proposition, SBERT has been used successfully in tasks where BERT had obtained good results [17, 18]. However, to the best of our knowledge, these applications do not yet include STS in Portuguese.

The goal of this work is to evaluate contextual embeddings generated by SBERT models for STS in Portuguese, which we investigate in two stages. First, we compare the performance of pre-trained SBERT models with the state-of-the-art BERT models for the ASSIN datasets [8]. In addition, we include other baseline models, such as the best-performing works assessed in the workshops and other multilingual contextual embeddings. Later, we evaluate the benefits of fine-tuning SBERT models for STS in Portuguese, also comparing them with all baseline and state-of-the-art models.

Results from the first part of our investigation showed that multilingual SBERT models are competitive, outperforming the best results of the ASSIN shared task. Even if at this stage fine-tuning was not considered, SBERT results were second only to the results achieved by the state-of-the-art BERT model [8]. In the second part of our work, results were improved with the fine-tuning, numerically surpassing the performance of the state-of-the-art model for the ASSIN 2 dataset, at a much lower computational cost. For the ASSIN datasets, even if the state-of-the-art results were not matched, the contextual embeddings generated by multilingual SBERT models with fine-tuning remained competitive and, as mentioned before, required a reduced computational cost. Finally, we discuss the impacts of fine-tuning and language variants with a qualitative assessment that demonstrates that results can be further improved in the future.

The remainder of this work is structured as follows. Section 2 briefly reviews preliminary concepts, namely contextual embeddings and the most relevant architectures used to produce them, and deep learning training approaches such as fine-tuning and knowledge distillation. Next, Section 3 defines the STS problem, details the ASSIN and ASSIN 2 shared tasks, and briefly discusses the state-of-the-art for these datasets. In Section 4, we detail the experimental setup adopted in this investigation, and discuss results in Sections 5 and 6. We conclude and discuss future work in Section 7.

2 Background

In this section, we briefly review the main preliminary concepts required to understand contextual embeddings, an important tool for problem solving in NLP. Initially, we discuss transfer learning [25], the training paradigm that motivated the proposal of embeddings in general. Next, we discuss contextual embeddings and the main algorithms adopted in this work, namely BERT [7] and Sentence-BERT [17]. Finally, we briefly detail knowledge distillation [10], the deep learning training paradigm that enables the multilingual training used to produce the models we adopt in this work.

2.1 Transfer learning

Transfer learning is a training paradigm that enables models fit to a general problem to be reused for a separate, and usually more specific problem [25]. When solving the new problem, two main strategies can be adopted.

Pre-trained models can be applied directly to the problem at hand when adjustments to match different input and/or output dimensionalities are feasible. Though pre-trained models are not expected to perform as well as problem-specific models, they are reusable across different problems.

Fine-tuned models are pre-trained models that are subject to additional training on a target specific problem. In detail, fine-tuning uses the pre-trained model as a starting point and runs additional training iterations on the new data.

Given the benefits of transfer learning, pre-trained models have become increasingly publicly available. Combined with fine-tuning, the same pre-trained model can become multiple specialized models without training from the ground up, saving both computing and time resources. In the context of NLP problems, this has motivated researchers to propose general models and make their pre-trained versions available, the most relevant example being the contextual and word embeddings we discuss next.

2.2 Contextual and word embeddings

A statistical language model is a probability distribution function of word sequences in a given language. Training language models is a very computationally intensive task, easily plagued by the curse of dimensionality, for two main reasons. First, the number of word co-occurrence is typically large. Second, the word sequences vary considerably across multiple text datasets, and even from training to test sets. These challenges motivated the distributed word representation [3], models in which both the distributed representation of each word and the probability function for the word sequence are learned simultaneously. Distributed representations can achieve better generalization since these models are able to attribute a high probability of occurrence to word sequences it had never seen before if similar sequences have been provided during training.

Popular word embeddings like Word2Vec [13], GloVe [14], and FastText [4] consider all sentences in which a given word appear to build a global vector representation for each word in the corpus. However, a word can have different meanings depending on the sentence and context it is in. To solve this problem, different contextual embeddings were proposed. Next, we discuss the two contextual approaches we use in this work, namely BERT and Sentence-BERT.

2.3 BERT

The *bidirectional encoder representations from Transformers* (BERT) architecture brought great advances to the NLP field, lately becoming the current state-of-the-art in different tasks [7]. BERT is a language model designed to create

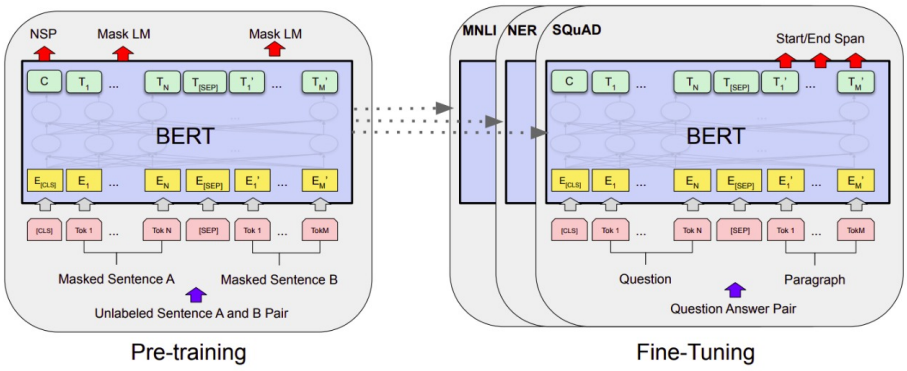


Fig. 1: The BERT architecture [7] for pre-training (left) or fine-tuning (right).

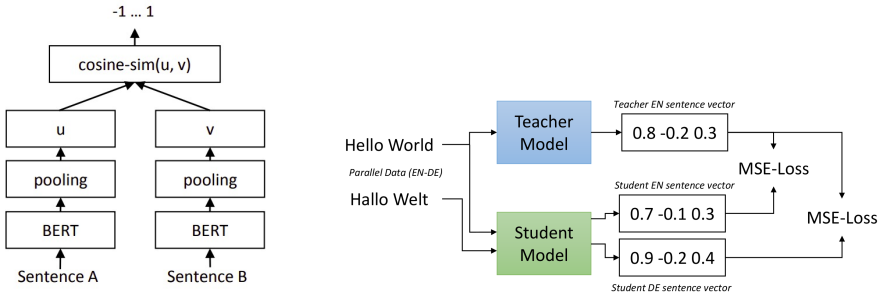
deep bidirectional vector representations from unsupervised texts. The bidirectional approach models context both to the right and to the left of each token in the input sequence. The BERT architecture can be seen on Figure 1, where the pre-trained scenario is seen on the left and the fine-tuning scenario is illustrated for different tasks on the right.

An improvement over other architectures is that BERT can be applied to tasks that take as input either individual sentences or pairs thereof. In the latter case, both sentences are separated by a special token dubbed [SEP]. Whether for individual or sentence pairs, the first token fed to the network is a special token dubbed [CLS], which serves as encoding for supervised tasks after training. Alternatively, a BERT embedding can be obtained by aggregating the three types of internal BERT embeddings, namely token, segment, and position.

2.4 Sentence-BERT (SBERT)

Although BERT has become the state-of-the-art for several NLP tasks, the algorithm faces scalability issues in tasks such as STS where there is the need to feed the network with sentence pairs. According to Reimers and Gurevych [17], even a moderate sized training corpus of 10000 sentences requires 50 million computational inferences, rendering BERT impractical for STS. As an alternative, SBERT was proposed as a more efficient approach to this type of tasks .

SBERT [17] is a siamese architecture in which the sentence pairs are fed to two parallel but connected network pathways to produce an estimate of the similarity between the sentences (Figure 2a). The first layers consist of regular BERT networks followed by a pooling operation to ensure embeddings produced for each sentence have the same vector size. The network outputs the cosine similarity between these embeddings. Since both BERT networks take a single sentence and their weights are shared, the number of computational inferences is much reduced [17].



(a) SBERT architecture [17] (b) Knowledge distillation [6].

Fig. 2: SBERT architecture (left) and multilingual knowledge distillation (right).

2.5 Knowledge distillation

Another concept relevant to this study is knowledge distillation, one of the deep learning training approaches where one model is trained to learn the knowledge encoded by another [10]. In this training approach, a *student* model attempts to emulate a *teacher* model with similar (or even higher) performance.

A knowledge distillation-based approach has been proposed by Reimers and Gurevych [17] to produce multilingual SBERT models and is illustrated in Figure 2b. The teacher model is an English pre-trained SBERT model and the student model is trained to produce multilingual embeddings. Training is structured as follows. First, the input for teacher and student models are different, since the original sentence fed to the teacher is enriched with parallel, translated versions for the student. Second, the objective function is to reduce the mean squared error losses comparing every student embedding produced with the original embedding produced by the teacher. Effectively, translated sentences are mapped to the same location in the Euclidean space of the original sentence.

Two other aspects of that work are further worth mentioning. First, the student models are trained with parallel data for 50 languages including Portuguese. Second, the student model is learned by a XLM-R algorithm [6], a *cross-lingual* version of BERT. In detail, authors claim XLM-R is better suited for multilingual learning than BERT given its ability to handle non-latin alphabets.

As reviewed in this section, the literature on contextual embeddings has matured over recent years. Still, models are usually developed and made available for NLP applications in English. In the next section, we discuss the STS problem in the context of the Portuguese language.

3 Semantic textual similarity in Portuguese

As previously discussed, semantic textual similarity (STS) is a relevant NLP application with international efforts to foster its research. STS can be defined as a regression task where one wants to numerically measure the similarity between two sentences. Higher scores indicate a stronger similarity between a sentence

pair. Among the main international efforts concerning STS, the *International Workshop on Semantic Evaluation* (SemEval [1]) is a series of workshops focused on semantic NLP problems. Each year, several shared tasks are hosted, in which high-quality datasets are used to benchmark submissions from participating research teams. Several similarity tasks have been defined at SemEval, including *cross-lingual* and *multilingual* STS [5].

The first dataset that included semantic textual similarity between pairs of sentences in Portuguese was ASSIN [9]. Later, ASSIN 2 [15] proposed a new dataset, based on sentences from the SICK-BR dataset [16]. Next, we describe each shared task, its associated datasets, and the best results observed for each.

ASSIN stands for *evaluation of semantic textual similarity and natural language inference* [9], a shared task organized at the PROPOR 2016 conference. The ASSIN datasets comprise (i) PT-PT, containing sentences in European Portuguese, and (ii) PT-BR, containing sentences in Brazilian Portuguese. Each dataset (PT-PT and PT-BR) has 5,000 sentences: 2,500 for training, 500 for validation, and 2,000 for testing. Sentences in each dataset were obtained from Google News, processed in three steps. First, vectorial space models [21] were used to select similar sentences from different documents. Second, these sentences were filtered manually, and noisy pairs were removed. Finally, filtered sentence pairs were annotated by four human judges. Scoring rules are describe below:

1. Sentences are completely different. They may talk about the same fact, but this cannot be determined without context.
2. Sentences refer to different facts and are not similar, but are about the same subject (e.g. a soccer match or elections).
3. The sentences present some similarity; can refer to the same fact or not.
4. Sentence content is very similar, but one (or both) has some exclusive information (e.g. a date or locality).
5. Sentences have practically the same meaning, possibly with a minimal difference (e.g. an adjective that does not change interpretation).

ASSIN 2 was the second edition of ASSIN [15], held at the STIL 2019 symposium in Brazil. The dataset is a translation and manual adaptation of the SICK dataset [12], the SICK-BR dataset [16]. In addition, ASSIN 2 also includes manually-generated sentence pairs, which were reviewed by human judges. Each pair of sentences in ASSIN 2 was annotated by at least four native Brazilian Portuguese speakers with linguistic training background. For STS, the final score was the average of the scores of each annotator. The final dataset has 10000 pairs of sentences: 6500 used for training, 500 for validation, and 3000 for testing.

Baseline results for the ASSIN and ASSIN 2 datasets are listed at the top of Table 1. Pearson correlation coefficient (to be maximized) is considered as the primary metric for the comparison, complemented by the mean squared error (MSE, to be minimized). The three topmost works are the best results obtained during the shared tasks, whereas the following two models were proposed in a work published after ASSIN 2 [8].[5] For brevity, the models we assess in this work are also given in Table 1, and are described in the next section.

[5] For brevity, other relevant works such as [19] comparing Word2Vec, FastText, ELMO, and BERT on ASSIN are not included as their results are surpassed by [8].

Table 1: Results using baseline, pre-trained, and fine-tuned models.

	ASSIN PT-BR		ASSIN PT-PT		ASSIN 2	
	Pearson	MSE	Pearson	MSE	Pearson	MSE
Baseline models						
Solo Queue (Workshop)	0.70	0.38	0.70	0.66	—	—
L2F/INESC-ID (Workshop)	—	—	0.73	0.61	—	—
IPR (Workshop)	—	—	—	—	0.82	0.52
ptBERT-Base$_{ft}$ [8]	0.83 ±0.00	0.25	0.85 ±0.00	0.47	0.84 ±0.01	0.50
ptBERT-Large$_{ft}$ [8]	**0.84 ±0.01**	**0.23**	**0.85** ±0.01	**0.40**	**0.84** ±0.01	**0.43**
BERTimbau-Base [20]	0.44	1.97	0.38	2.46	0.62	1.78
BERTimbau-Large [20]	0.55	3.32	0.52	4.16	0.27	2.49
Pre-trained models						
BERTimbau-Base (SBERT)	0.57	1.60	0.53	2.22	0.68	1.24
BERTimbau-Large (SBERT)	0.58	2.69	0.49	3.57	0.71	1.92
⟨XLM-R, SBERT, NLI+STS⟩	0.67	0.65	0.71	0.68	0.79	0.63
⟨XLM-R, SBERT, Paraphrases⟩	0.71	**0.42**	0.74	0.65	0.79	**0.50**
⟨XLM-R, SBERT, Cross-EN-DE⟩	**0.73**	0.59	**0.77**	**0.48**	**0.80**	0.64
Fine-tuned models						
BERTimbau-Base (SBERT)$_{ft}$	0.76 ±0.01	0.35	0.74 ±0.00	0.61	0.83 ±0.00	0.52
BERTimbau-Large (SBERT)$_{ft}$	0.78 ±0.01	0.32	0.76 ±0.01	0.68	0.84±0.00	0.53
⟨XLM-R, SBERT, NLI+STS⟩$_{ft}$	0.76 ±0.00	0.35	0.74 ±0.00	0.54	0.84 ±0.00	0.42
⟨XLM-R, SBERT, Paraphrases⟩$_{ft}$	**0.79** ±0.00	**0.30**	**0.77** ±0.00	**0.52**	**0.85** ±0.00	0.43
⟨XLM-R, SBERT, Cross-EN-DE⟩$_{ft}$	0.78 ±0.00	0.31	**0.77** ±0.00	0.55	**0.85** ±0.00	**0.41**

Baseline results. The work of Fialho, Coheur, and Quaresma [8] considerably advanced the state-of-the-art for both ASSIN PT-PT and ASSIN PT-BR, and to a lesser extent also for ASSIN 2. Specifically, the authors experimentally assessed BERT models for STS and NLI in Portuguese using the datasets ASSIN, ASSIN 2, and SICK-BR. For STS, two approaches were considered: BERT as a regressor or BERT embeddings as input for regression algorithms. Three models fine-tuned for STS were evaluated, namely (i) one multilingual and (ii) two monolingual versions pre-trained in Brazilian Portuguese differing in size (ptBERT-Base$_{ft}$ and ptBERT-Large$_{ft}$). ptBERT-Large$_{ft}$ used as a regressor achieved the best results.

As discussed in this section, ASSIN and ASSIN 2 are relevant Portuguese STS datasets for which BERT models are the state-of-the-art. Yet, the computational overhead incurred for BERT training limits its applicability to problems with a moderate size sentence collection. In the next section, we discuss how we employ and assess SBERT models on the Portuguese STS datasets reviewed here.

4 Experimental assessment

SBERT models have been successfully employed on NLP applications where BERT models had state-of-the-art performance [17]. Furthermore, multilingual pre-trained SBERT models have been made available in an open source SBERT repository (https://www.sbert.net). In this section, we describe our experimental

analysis in which we assess the performance of SBERT models with and without fine-tuning for STS in Portuguese. Initially, we detail the models we consider, following the order given in Table 1.[6] Later, we discuss the two stages of our experiments, differing as to whether fine-tuning is employed.

4.1 Models assessed

As listed in Table 1, the top-most five models we consider as baseline are (i) the best-performing algorithms of the ASSIN and ASSIN 2 shared tasks, and (ii) the state-of-the-art BERT models [8], already described in the previous section. In addition, we also include as baseline two Portuguese pre-trained BERT models called BERTimbau [20]. Remaining models given in Table 1 are used within the SBERT architecture. The first two reuse BERTimbau weights, given the compatibility between BERT and SBERT. The remaining three models are multilingual pre-trained models open sourced by SBERT proponents.

We choose to employ pre-trained models instead of training models from scratch for two major reasons. First, the computational cost required for training deep learning NLP models from scratch is considerable. Second, the pre-trained models have often been trained on very rich corpora, whether in Portuguese or multilingual. Below, we describe the BERTimbau and multilingual models.

BERTimbau models were trained by *Neuralmind* and are available in two sizes, namely BERTimbau-Base and BERTimbau-Large. Pre-training BERT requires a vocabulary and a training corpus. *Neuralmind* produced a vocabulary from the Brazilian Portuguese version of the Wikipedia, and used the *brWac* corpus [22]. Pre-training followed the original BERT paper [7]. BERTimbau-Base and BERTimbau-Large were respectively warm-started with the *multilingual BERT-Base* and *English BERT-Large* checkpoints.

SBERT models pre-trained in Portuguese are not openly available. Instead, we adopt two sets of models, given in Table 1 on the five rows that follow baseline models. The first two models, labeled BERTimbau-Base (SBERT) and BERTimbau-Large (SBERT), reuse the trained weights provided by BERTimbau models. The remaining three models are multilingual models open-sourced by SBERT proponents, having been trained using the knowledge distillation process. Among the models available, we used the ones for which the student-model was learned by the XLM-R algorithm [6]. Different teacher-models were used, described below under the notation ⟨student, teacher, task⟩, indicating the student-model and the task for which the teacher-model was originally trained.

- ⟨XLM-R, SBERT, NLI+STS⟩: XLM-R learned a multilingual model from an SBERT model fine-tuned for NLI and STS in English.
- ⟨XLM-R, SBERT, Paraphrases⟩: XLM-R learned a multilingual model from an SBERT model trained on an English paraphases dataset.
- ⟨XLM-R, SBERT, Cross-EN-DE⟩: fine-tuned ⟨XLM-R, SBERT, Paraphrases⟩ model for the STSbenchmark [5], with English and/or German sentences.

[6] Though models based on other relevant architectures such as the multilingual universal sentence encoder [23] were available at the SBERT repository, we did not include them in our work due to the lack of training setup details.

Training multilingual versions used a range of datasets, and so contextual embeddings produced by SBERT were generalized by XLM-R for 50 different languages. Further information about their training is available in the original paper [18].

4.2 Experimental setup

As previously discussed, we split our experiments in two parts to isolate the effect of fine-tuning. This is reflected in Table 1, where we label results obtained from a given model after fine-tuning with the subscript *ft*. Each part of the investigation is further detailed next.

Experiments without fine-tuning. In the first part of our experiments, we employed the pre-trained models on ASSIN and ASSIN 2 datasets without fine-tuning. Concretely, labels provided by the datasets were used only to validate the results. For the baseline BERTimbau models (BERTimbau-Base and BERTimbau-Large), each input sentence was fed individually to the model and the corresponding embedding was obtained from the [CLS] token. The similarity score for a pair of sentences was calculated as the cosine distance between the sentence embeddings produced. Then, we compare the cosine similarity directly with the labels of ASSIN and ASSIN 2 using Pearson correlation. Though cosine distance ranges differs from the STS score range, normalization and scaling is not necessary since Pearson correlation only considers the linear relationship between the variables.

Experiments with fine-tuning. In the second part of our experimental analysis, we assessed the impact in performance when pre-trained models are fine-tuned for the ASSIN and ASSIN 2 datasets. In detail, we use the available validation labels to fine-tune the pre-trained models. The developers of SBERT made available a standard fine-tuning script, with recommended hyperparameters we adopted. For each dataset and model, we ran fine-tuning ten times and report mean and standard deviation results. Regarding baseline BERTimbau results, fine-tuning would be computationally unfeasible in the context of this work. Yet, we remark that the state-of-the-art models [8] are ptBERT-Base and ptBERT-Large networks that have been fine-tuned for the ASSIN and ASSIN 2 datasets. As such, these models serve as reference for BERT results, though mean and standard deviation have been reported only for five repetitions.

Given the large number of models we consider, our complete set of results is provided as supplementary material [2]. In the next sections, we discuss the most relevant insights we observe in our assessment.

5 Results

As previously discussed, contextual embeddings are usually produced from pre-trained or fine-tuned language models. In our setup, we have isolated experiments without and with fine-tuning for STS in Portuguese. Below, we discuss the most relevant insights observed for each part of our investigation.

5.1 Pre-trained models without fine-tuning for STS in Portuguese

In order of appearance in Table 1, literature results are given at the top, as previously detailed. The following two models are also taken as baseline. In detail, BERTimbau-Base and BERTimbau-Large are the BERTimbau models that have been pre-trained in Portuguese. Remaining models are SBERT approaches. Given the compatibility between BERT and SBERT, we include SBERT results where weights were obtained from BERTimbau models, which we refer to as ptBERT-Base (SBERT) and ptBERT-Large (SBERT). Finally, Table 1 lists the multilingual SBERT models, namely ⟨XLM-R, SBERT, NLI+STS⟩, ⟨XLM-R, SBERT, Paraphrases⟩ and ⟨XLM-R, SBERT, Cross-EN-DE⟩. Next, we discuss the main insights observed in this part of our investigation.

Portuguese pre-trained models did not produce competitive results w.r.t. workshop or literature results. A direct comparison between models that use BERTimbau weights showed that SBERT models in general achieved better results than their BERT counterparts. This result is interesting considering that the only difference between them is that embeddings from BERT models are obtained from the [CLS] special token, whereas SBERT is average pooling all BERT output embeddings. Another important insight is that not always the Large version outperformed its Base counterpart. This is likely explained by the checkpoints used by *Neuralmind* for BERTimbau model training. As discussed in Section 4, a multilingual BERT checkpoint was used to warm start Base training, whereas Large models were warm-started with an English BERT checkpoint.

Multilingual pre-trained models produced competitive results w.r.t. results from the participating teams of the shared tasks. This is true for nearly all datasets and multilingual models considered, with two major exceptions. First, the ⟨XLM-R, SBERT, NLI+STS⟩ model is unable to match workshop results for any of the datasets considered, though its teacher-model had been originally trained for NLI and STS tasks. Second, no model is able to match the performance of ASSIN 2 workshop results. Among the SBERT versions, the one that shows the most consistent performance is the one fine-tuned for Cross-EN-DE STS. This indicates that fine-tuning for multilingual tasks is helpful, as expected. Yet, without a specific fine-tuning for STS in Portuguese, even this model cannot match state-of-the-art results [8]. Indeed, in this part of the investigation all SBERT results fall short of the state-of-the-art for all datasets considered.

5.2 Pre-trained models with fine-tuning

Results obtained with fine-tuned models are displayed at the bottom part of Table 1. Models are given in the same order as the pre-trained models. Below, we discuss the main insights from this analysis.

Benefits of fine-tuning. The performance gains provided by fine-tuning are considerable for all algorithms. In fact, all models now (numerically) outperform the best results achieved during the ASSIN and ASSIN 2 tasks. To evidence the performance gains provided by fine-tuning, the boxplots in Figure 3 show the average performance of all SBERT models with and without fine-tuning,

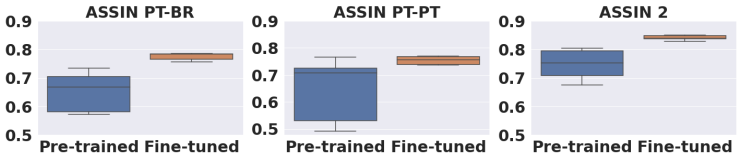


Fig. 3: Performances of pre-trained models with and without fine-tuning.

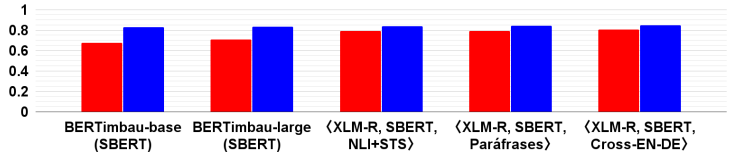


Fig. 4: Homogeneity of results with (blue) and without (red) fine-tuning

grouped by dataset. In addition to improving average performance, fine-tuning also reduces variability to a significant extent, especially on ASSIN PT-PT.

Comparison with the state-of-the-art. Results differ as a function of shared task. Multilingual SBERT models numerically surpass the state-of-the-art results according to Pearson correlation and/or MSE on ASSIN 2. Remaining models also get closer to the state-of-the-art for this task, confirming the expected benefits from SBERT and fine-tuning. For ASSIN PT-BR and ASSIN PT-PT, the best results showed by SBERT models are not yet competitive.

Homogeneity of results. In the absence of fine-tuning (*pre-trained* section in Table 1) results were very contrasting among different models. Conversely, all fine-tuned models have produced results that are much closer to each other. We illustrate this increased homogeneity for ASSIN 2 results in Figure 4, where average pre-trained model performance is given in red and average fine-tuned model performance is given in blue, grouped by model and sorted by fine-tuned performance. For this dataset, even BERTimbau models lead to a competitive performance after fine-tuning.

Overall, results discussed in this section confirmed the benefits of adopting SBERT and fine-tuning for STS in Portuguese. For the ASSIN 2 dataset, multilingual models even numerically surpass the current state-of-the-art. In the next section, we further investigate results to understand these benefits.

6 Further analysis

To better understand some of the insights observed in the previous section, we initially conduct a qualitative assessment, regarding both (i) fine-tuning and (ii) language variant effects. Moreover, we run additional experiments to assess the benefits of multiple language variants in the fine-tuning setup.

Table 2: ASSIN 2 sentence pair examples with their scores illustrating changes in cosine similarity before and after fine-tuning ⟨XLM-R, SBERT, Paraphrases⟩.

Sentence pairs	**Score**	**Before**	**After**
Examples of sentence similarities that **improved** after fine-tuning			
S1: "A senhora está pegando o canguru." S2: "Um canguru está pegando o bebê da senhora."	2.0	3.9	2.0
S1: "Uma senhora asiática está colocando maquiagem." S2: "A senhora está se maquiando."	4.0	2.2	4.1
S1: "A senhora está limpando um camarão" S2: "Alguém está limpando um animal"	4.0	2.6	4.1
Examples of sentence similarities that **deteriorated** after fine-tuning			
S1: "Um homem está colocando um dispositivo eletrônico." S2: "O homem está tirando uma foto dele mesmo e de outro cara."	1.9	1.9	3.4
S1: "Um homem está dançando." S2: "Não tem nenhuma mulher se exercitando."	1.2	1.1	2.4
S1: "A senhora está andando de elefante." S2: "O elefante está sendo montado pela senhora."	3.8	3.5	4.8

6.1 Fine-tuning effects

We select a few sentence pairs to illustrate fine-tuning effects on ⟨XLM-R, SBERT, Paraphrases⟩ results, the SBERT model that performed best in our experiments. Table 2 gives six sentence pair examples with their expected scores and cosine similarities before and after fine-tuning for STS in Portuguese. For the first three pairs, results were improved by fine-tuning. Indeed, fine-tuned results become very close to the expected score, whereas pre-trained results were very off.

By contrast, results obtained with fine-tuning for the latter three sentence pairs in Table 2 are worse than when fine-tuning is not adopted. In the first deterioration example, we believe the model misinterprets the terms "*colocando*" (putting) e "*tirando*" (taking), which can be used as antonyms but not in the context "*tirando uma foto*" (taking a picture). In the second example, we believe the model misinterprets a negation quantifier "*Não tem nenhuma mulher*" (no woman) with an opposition adverb (not a woman), which renders both sentences less semantically different. Future analysis could also investigate the impact of double negatives in this case, as multilingual models need to deal with such differences between languages. Finally, in the last example the only differences between sentences regard (i) active versus passive sentence form, and (ii) whether "*montando*" refers to getting on or riding an animal. Since both interpretations are possible, we believe that the ASSIN 2 example could be considered noisy.

6.2 Language variant effects

With the exception of the state-of-the-art, all algorithms assessed for ASSIN datasets in the literature and also here are unable to perform their best for the European and Brazilian Portuguese variants at the same time. We illustrate this with sentence pairs given in Table 3. The two top-most sentence pairs are obtained from ASSIN PT-BR, and the two bottom-most from ASSIN PT-PT.

Table 3: ASSIN PT-BR (top) and PT-PT (bottom) pairs with their scores illustrating fine-tuning ⟨XLM-R, SBERT, Paraphrases⟩ for different language variants.

Sentence pairs	Score	BR	PT
ASSIN PT-BR examples			
"O show está previsto para começar às 17h." "A atração acontece a partir das 12h."	1.8	1.5	3.0
"Já para 2016, a previsão dos economistas recuou de 5,6% para 5,51%." "A mediana das estimativas passou de 5,76% para 5,71%."	2.2	2.2	3.5
ASSIN PT-PT examples			
"Noruega e Roménia defrontam-se ainda esta quinta-feira no Pavilhão Municipal da Póvoa de Varzim. "Os bilhetes para os jogos que decorrerão no Pavilhão Muncipal da Póvoa do Varzim são gratuitos."	2.0	1.5	2.5
"A companhia aérea brasileira Azul lança a 4 de maio o seu primeiro voo regular para a Europa." "O voo inaugural da Azul acontece já a 4 de maio."	4.0	3.1	3.9

Sentences selected in European Portuguese respectively illustrate vocabulary and grammatical differences between the language variants. Beside expected scores, similarities computed by ⟨XLM-R, SBERT, Paraphrases⟩ when fine-tuned for ASSIN PT-BR and ASSIN PT-PT are also given. Results for the model fine-tuned for ASSIN PT-BR are better on that dataset. The symmetrical situation is observed for the ASSIN PT-PT dataset.

The better performance of models fine-tuned for the given language variant is expected. We then conduct additional experiments to understand the impact of fine-tuning ⟨XLM-R, SBERT, Paraphrases⟩ for the three ASSIN and ASSIN 2 datasets altogether. We report mean values for ten repetitions, and provide standard deviations as supplementary material [2] given the very low variability observed in the results. Results differ as a function of the target testing dataset. For ASSIN PT-BR and ASSIN 2, results are not changed, except for MSE on the latter, reduced from 0.43 to 0.41. Regarding ASSIN PT-PT, results are strongly improved, with Pearson coefficient correlation increasing from 0.77 to 0.81 and MSE reducing from 0.52 to 0.49.

The qualitative assessment conducted in this section helped further understand the benefits of fine-tuning and the potential impacts of different language variants. Even if improved results did not match the state-of-the-art for the ASSIN PT-PT dataset, they shed light into promising future work directions.

7 Conclusions

Semantic textual similarity (STS [11]) is a core research problem in natural language processings (NLP) studies. This is evidenced by the (i) international efforts promoting STS shared tasks [12, 5] and (ii) the fact that several novel contextual embedddings recently proposed have been benchmarked on this problem [17]. In the context of the Portuguese language and its variants, ASSIN and ASSIN 2 are the most relevant shared tasks identified [9, 15], for which BERT embeddings comprise the current state-of-the-art [8]. In this work, we have investigated the

application of the Sentence-BERT (SBERT [17]) architecture to STS in Portuguese. SBERT addresses scalability issues in BERT networks, and has been successfully applied and open-source also for multilingual applications [18].

Our contributions were two-fold. In the first part of this work, we assessed Portuguese and multilingual SBERT models, demonstrating that the latter present competitive performance w.r.t. shared task results even without fine-tuning. This can be very useful to data professionals who need to address this problem in unsupervised learning scenarios, as typical in real-world applications. In the second part, we demonstrated that fine tuning SBERT led to improvements in all previous results. The multilingual models of SBERT once again stood out and this time surpassed the state-of-the-art for the ASSIN 2 dataset.

Further analysis indicated promising future work possibilities. In detail, we have observed situations where fine-tuning worsened results, an indication that pre-training SBERT models in Portuguese could lead to yet better results. More importantly, we have discussed the impact of language variants, which we believe should be taken into consideration either for pre-training or fine-tuning. We evidenced this possibility with additional experiments that further improved the performance on European Portuguese for the best-performing SBERT multilingual model.

The contributions discussed in this paper further highlight that other contextual *embeddings* such as XLNet [24] need to be assessed in the context of STS in Portuguese. In this sense, we believe our work is instrumental to motivate more research groups addressing NLP problems in Portuguese. This is the purpose of shared tasks, especially given the computational cost of the experimental campaigns involved in deep learning assessment for NLP applications. Even more interesting is stirring the interest of researchers proposing novel NLP models to the Portuguese language, given that the differences between Portuguese variants impacts on the results for most models considered.

References

1. Agirre, E., Diab, M., Cer, D., Gonzalez-Agirre, A.: SemEval-2012 task 6: A pilot on semantic textual similarity. In: SemEval. p. 385–393. ACL, USA (2012)
2. Andrade, J., Bezerra, L.C.T., Cardoso-Silva, J.: Comparing contextual embeddings for semantic textual similarity in portuguese (supplementary material). (2021), https://github.com/andradejunior/bracis-2021-supp-material
3. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
4. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Trans. of the ACL **5**, 135–146 (2017)
5. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: SemEval. pp. 1–14. ACL, Vancouver, Canada (2017)
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: ACL. pp. 8440–8451. ACL (2020)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL. pp. 4171–4186. ACL, Minneapolis, Minnesota (2019)

8. Fialho, P., Coheur, L., Quaresma, P.: Benchmarking natural language inference and semantic textual similarity for portuguese. Information **11**, 484 (2020)
9. Fonseca, E.R., Borges dos Santos, L., Criscuolo, M., Aluísio, S.M.: Visão geral da avaliação de similaridade semântica e inferência textual. Linguamática **8**(2), 3–13 (2016)
10. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. Intern. J. of Comput. Vis. **129**(6), 1789–1819 (2021)
11. Jurafsky, D., Martin, J.H.: Speech and Language Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2 edn. (2009)
12. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., Zamparelli, R.: SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In: SemEval. pp. 1–8. ACL, Dublin, Ireland (2014)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NeurIPS. p. 3111–3119. Curran Associates Inc., Red Hook, NY, USA (2013)
14. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: EMNLP. pp. 1532–1543. ACL, Doha, Qatar (2014)
15. Real, L., Fonseca, E., Gonçalo Oliveira, H.: The ASSIN 2 shared task: A quick overview. In: Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., Gonçalves, T. (eds.) PROPOR. pp. 406–412. Springer International Publishing, Cham (2020)
16. Real, L., Rodrigues, A., Vieira, A., Albiero, B., Thalenberg, B., Guide, B., Silva, C., Lima, G., Câmara, I., Stanojević, M., Souza, R., De Paiva, V.: SICK-BR: A Portuguese corpus for inference. In: PROPOR. pp. 303–312. Springer (2018)
17. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: EMNLP. pp. 3973–3983. ACL (2019)
18. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: EMNLP. pp. 4512–4525. ACL (2020)
19. Rodrigues, R.C., Rodrigues, J., de Castro, P.V.Q., da Silva, N.F.F., Soares, A.: Portuguese language models and word embeddings: Evaluating on semantic similarity tasks. In: Quaresma, P., Vieira, R., Aluísio, S., Moniz, H., Batista, F., Gonçalves, T. (eds.) Computational Processing of the Portuguese Language. pp. 239–248. Springer International Publishing, Cham (2020)
20. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: pretrained BERT models for Brazilian Portuguese. In: BRACIS. pp. 403–417. Springer (2020)
21. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. J. Artif. Int. Res. **37**(1), 141–188 (Jan 2010)
22. Wagner Filho, J.A., Wilkens, R., Idiart, M., Villavicencio, A.: The brWaC corpus: A new open resource for Brazilian Portuguese. In: LREC. ELRA, Miyazaki, Japan (2018)
23. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.h., Strope, B., Kurzweil, R.: Multilingual universal sentence encoder for semantic retrieval. In: ACL: System Demonstrations. pp. 87–94. ACL, Online (Jul 2020)
24. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: XLNet: Generalized autoregressive pretraining for language understanding. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) NeurIPS. vol. 32. Curran Associates, Inc. (2019)
25. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) NIPS. vol. 27. Curran Associates, Inc. (2014)